F. 6. *Schema.org*: la catalogación revisitada

Francisco-Javier García-Marco

16 diciembre 2012

García-Marco, Francisco-Javier (2013). "*Schema.org*: la catalogación revisitada". *Anuario ThinkEPI*, v. 7, pp. 169-172.

Resumen: *Schema.org* supone una iniciativa y un avance importante en dos direcciones: por un lado, la incorporación de la world wide web a la normalidad en el ámbito de la recuperación de información, en la que la categorización en campos y el control del vocabulario se usan para mejorar la llamada o exhaustividad y la precisión, junto con herramientas lógicas, vectoriales y probabilísticas; y, por otro, la democratización de la web semántica o, si se quiere, el lanzamiento de una web semántica fácilmente incorporable por cualquier webmaster, que ahora puede ser casi cualquier persona con acceso a internet.

Palabras clave: *Schema.org*, Microformatos, Recuperación de la información, Web semántica.

Title: *Schema.org*: cataloguing revisited

Abstract: *Schema.org* is an important initiative that advances in two directions: on one hand, towards information retrieval normalcy on the world wide web, where fields and vocabulary control are used to improve recall and precision in addition to other logical, vector and probabilistic tools; and, on the other hand, the democratization of the semantic web or, in other words, the launch of a semantic web that can be easily adopted by any willing web master, which now can be almost any person with access to the internet.

Keywords: *Schema.org*, Microformats, Information retrieval, Semantic web.

Introducción

El 2 junio de 2011 se produjo una noticia importante para los interesados en el tratamiento controlado de la información, hasta hace poco denominado catalogación o descripción documental e indización, y hoy referido cada vez más frecuentemente como asignación de metadatos.

En tal fecha *Google*, *Microsoft* y *Yahoo!* comunicaron que habían acordado trabajar juntos para convencer a los administradores de webs para que estructuraran sus páginas según esquemas comunes, de manera que la recuperación de su información fuera más relevante y exhaustiva. Se ponía en marcha la iniciativa *Schema.org* (2012).

Los esquemas propuestos en *Schema.org* son microformatos o microplantillas de catalogación para diversos tipos de información que se expresan en *RDF Schema* —por lo que se integran en la web semántica- y son soportados por los buscadores más importantes del mundo, incluyendo también al buscador ruso *Yandex* desde noviembre de 2011.

Los grandes buscadores no han propuesto esta aproximación a la catalogación por hacer un favor a los que creemos en el control de datos dentro de esquemas de representación normalizados, sino porque sus sistemas están dejando de ser eficaces para proporcionar una conexión relevante para el usuario entre lo que busca y los anuncios que le pueden interesar. Una desconexión que amenaza sus resultados corporativos.

También están interesados en este enfoque porque quieren desarrollar –de hecho lo hacen desde hace algunos años– servicios de agregación y comparación de todo tipo que requieren que los datos estén etiquetados para que funcionen automáticamente. Ejemplos sencillos y generalizados con gran valor económico por su potencial publicitario son los comparadores de precios, como *Google shopping*.

Como etiquetar es caro y se requiere personal especializado para hacerlo bien —por eso triunfan en las alternativas especializadas[1]—, los grandes buscadores, lógicamente, desean que

la catalogación sea hecha directamente por los proveedores de la información.

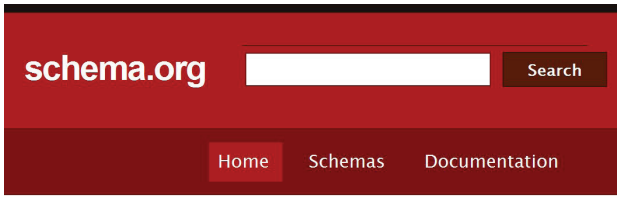
Por ello, un factor importante para el triunfo de esta iniciativa es que los microesquemas sean realmente adoptados por los responsables de los sitios web. Eso requiere tanto formación como, sobre todo, incentivos. La parte de formación se apoya, en primer lugar, en su gran sencillez, centrada en el tipo de información que se debe esquematizar. Pero lo que se espera que sea más importante es el incentivo para los administradores de webs, a saber, la mejora en el posicionamiento de sus páginas, cada vez más sepultadas en el océano de información en el que se ha convertido la *world wide web*.

"La Web está dotada de fuertes estructuras jerárquicas que complementan a las asociativas"

La web se está transformando en un sistema de recuperación normal

Por encima de los detalles técnicos y políticos de la iniciativa, es importante que las mayores empresas de internet redescubran la catalogación como instrumento para una conseguir una recuperación más precisa y exhaustiva.

En este sentido, *Schema.org* es radicalmente distinto del otro pilar del despegue de la catalogación en internet, *Dublin Core* (DC). Este surgió del ámbito de la biblioteconomía –apadrinado por *OCLC*–, mientras que *Schema.org* es una iniciativa que ha nacido en el campo de los motores de búsqueda, en definitiva, en el campo de la recuperación de la información.



Una preocupación inmediata es cómo se coordina esta iniciativa con otras de semejante pretensión de universalidad, especialmente la *Dublin Core Metadata Initiative* (*DCMI*). *DCMI* ha creado un *DCMI Schema.org alignment task group* (2012), cuyo objetivo es precisamente elaborar mapeos entre ambos estándares. Es significativo que sea *DCMI* quien quiera mapear, y *Schema.org* quien lance la iniciativa de forma independiente; resulta revelador de cuál es el equilibro de fuerzas.

Lo importante es que aproximaciones documentales que se creían superadas en el entorno web están siendo recuperadas. Lo cierto es que esto no ha pasado por primera vez: se reproduce por lo menos en la historia de la indización y recuperación postcoordinada, y en la del hipertexto.

La invención de las primeras máquinas ordenadoras y extractoras provocó la revolución de la indización postcoordinada extractiva automática, que explotó con la invención de la computadora electrónica, y luego fue corregida hasta desembocar en el necesario complemento de la indización postcoordinada controlada y el tesauro. Igualmente, el automatismo y la inmediatez de la navegación hipertextual sugirió un mundo asociacionista donde la lectura no estuviera basada en estructuras jerárquicas intratextuales o intertextuales; pero la Web hoy en día está dotada de fuertes estructuras jerárquicas que complementan a las asociativas, y el mapa del sitio y el menú jerárquico –las taxonomías– son parte inevitable del modo estándar de publicar información en la world wide web.

"*Schema.org* es radicamente distinto del otro pilar del despegue de la catalogación en internet: Dublin Core"

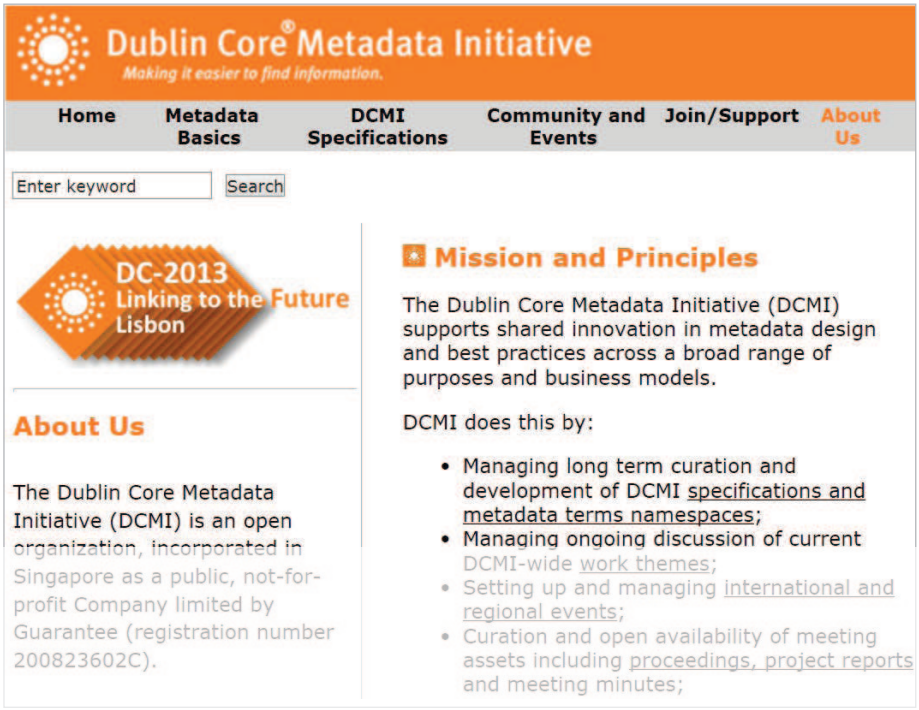
El hipertexto se planteó como alternativa a la lectura estructurada; y la búsqueda por palabras clave se impuso inicialmente a la recuperación dentro de esquemas organizados; pero ambas cosas no podían durar con exclusión de la otra parte de la realidad. Deslumbrados por el asociacionismo conceptual y la potencia de los índices automatizados, sus defensores dieron la espalda inicialmente a dos realidades que son psicológicas, no tecnológicas: por un lado, la jerarquización de conceptos es crítica para el funcionamiento de la memoria a largo plazo; y, por el otro, las estructuras de conceptos constituyen los espacios que hacen a la información semánticamente "navegable"[2].

La Web es otro ejemplo de redescubrimien-

to de los principios clásicos del tratamiento de la información[3]. Las primeras arañas se basaban en índices inversos de palabras extraídas y en operadores de búsqueda. La primera gran revolución se basó en la aplicación de técnicas documentales a la web: *Google* dotó de estructura a sus índices gracias al concepto de popularidad, inspirado en los índices de citas de **Eugene Garfield**. Lo mismo que los índices de palabras clave y la recuperación extractiva, los de citas son totalmente automatizables, dentro de parámetros tolerables de error.

Los buscadores han ido incorporando otras herramientas de control de vocabulario como la corrección ortográfica o los anillos de sinónimos, para evitar las anomalías que se producen en el proceso de emparejamiento de las preguntas de los usuarios y los contenidos de las bases de datos de los buscadores. Ahora se pretende una asignación de metadatos sistemática, ligada a la mejora del posicionamiento.

En el horizonte se atisba una iniciativa semejante en el campo de la Web Semántica con organización del conocimiento mediante ontologías. Pero estas tareas requieren inevitablemente de un mayor concurso humano. La Web ha ido paso a paso recorriendo el camino de reincorporación a los principios clásicos del procesamiento de la documentación que tuvo que realizar en su día la recuperación postcoordinada, incorporando las posibilidades de automatización reales a un paradigma de recuperación más amplio.



http://dublincore.org

A modo de conclusión: ¿reinvención de la rueda o recapitulación ontogénica de la filogenia informacional?

En los pocos meses que la iniciativa lleva en marcha, el número de esquemas[4] y —lo más revelador- su complejidad ha crecido a gran velocidad. Podemos predecir con bastante seguridad que sucederá lo mismo que con la catalogación bibliográfica o con los esquemas de descripción de documentos: su complejidad crecerá para acomodarse a las necesidades más exigentes, aunque será necesario preservar un conjunto mínimo de datos fácilmente comprensibles para que "todos" puedan catalogar sus documentos al menos de forma sencilla.

Al final, cada cual etiquetará equilibrando sus necesidades y sus recursos. Las operaciones críticas que se apoyan en ingresos importantes producirán una catalogación detallada; las que no, muy simple o inexistente; y entre medio, un amplio continuo.

"La Web es otro ejemplo de redescubrimiento de los principios clásicos del tratamiento de la información"

Es fascinante observar cómo los grandes motores de búsqueda dan un paso más hacia su "catalogización" forzados por la pérdida de relevancia y exhaustividad que el crecimiento exponencial de sus bases de datos ha provocado. Se trata de una reproducción del proceso que se produjo en la cultura del papel con la explosión de las publicaciones y su concentración en organizaciones —bibliotecas- cada vez mayores. Una reproducción, eso sí, acelerada, al trepidante ritmo de la tecnología informática, que recuerda la recapitulación que se produce en el desarrollo fetal de la evolución filogenética de la especie.

Los grandes directorios y catálogos de internet murieron; y con ellos parecía que se enterraban la catalogación y la clasificación en internet; por otra parte, el futuro de la información y la documentación. Recordemos el primer *Yahoo!*. Pero la necesidad seguía viva e, ignorada, ha terminado por hablar a gritos. Sin embargo, algo ha cam-

biado: la labor no se realizará centralizadamente por ahora en los grandes servicios de búsqueda de internet. Se trata ahora de que cataloguen los productores, no los agregadores. El que está en la parte de abajo de la pirámide trabaja; el que está en la parte de arriba, dirige y recoge. O, dicho de forma más neutra, que cada uno se centre en lo suyo. ¡Bienvenidos al *cataloguing in publication* (*CiP*) de internet!

Pero, en fin, ¿no buscamos trabajo? Pues parece que viene abundante. Cualquiera que mire desapasionadamente el esquema de la web semántica propuesto por **Tim Berners-Lee** (2001) o sus sucesivas reinterpretaciones, puede ver que reinserta a los bibliotecarios, documentalistas y archiveros dentro del gran proyecto de internet de una manera nueva pero a la vez asombrosamente clásica. Mientras siga habiendo energía abundante en el mundo para que funcione la Red, nuestra disciplina queda cada vez más dividida en dos grandes campos: los que trabajarán en su parte museística y de nicho —ligada a la preservación, el acceso y el comercio de la información en papel, celuloide y otros formatos físicos- y los que trabajarán en el nuevo espacio de información digital que, por otra parte, se "bibliotecariza" a pasos agigantados.

"Se trata ahora de que cataloguen los productores, no los agregadores"

Pero tampoco echemos las campanas al vuelo, porque con tanto cambio es difícil asegurar un coto cerrado. Aproximaciones como *Schema.org* se orientan sobre todo a promover también una catalogación "popular", al alcance de todos, motivada por un posicionamiento mejor en internet.

Enfoques como *Schema.org* no reconocen barreras profesionales ni de otro tipo. Es difícil, pues, que esto se convierta en monopolio profesional. Nuestro nicho seguirá estando, probablemente, en los proyectos de complejidad media, que exigen mantenimiento constante y que están ligados a un sustrato económico estable, basado en la producción de rentas o en la atracción de subvenciones suficientes.

A nivel más general, y saliendo de la perspectiva de nuestro nicho profesional —precioso ombligo, pero ombligo al fin y al cabo-, *Schema.org* supone una iniciativa y un avance importante en dos grandes direcciones: por un lado, la incorporación de la *world wide web* a la normalidad en el ámbito de la recuperación de información; y, por el otro, la democratización de la web semántica o, si se quiere, el lanzamiento de una web semántica fácilmente incorporable por cualquier webmaster, que ahora puede ser casi cualquier persona con acceso a internet que se tome un poco de molestia.

"Aproximaciones documentales que se creían superadas en el entorno Web están siendo recuperadas"

Notas

1. Especializadas no sólo en cuanto a las generalidades del tratamiento informacional, sino también sobre el campo específico que se trate, sea turismo, medicina o equipamiento informático.

2. Lógicamente, el otro extremo es igual de malo: que las jerarquías sean inflexibles, descarnadas y obsoletas; o la lectura rígida, encorsetada y descontextualizada. Pero la tecnología, especialmente, la web semántica, hace posible abordar también estos problemas.

3. ¿Son clásicos porque son principios, o se piensan como principios porque son clásicos? La recurrencia de su descubrimiento en los campos de la recuperación extractiva, el hipertexto y los motores de búsqueda proporciona una cierta evidencia a favor de la primera opción, aunque la recuperación —a diferencia de los procesos bibliológicos- de la información -que sea algo dado- es una técnica surgida de la invención humana, y puede ser siempre cuestionada por un nuevo paradigma.

4. El catálogo de tipos ha ido evolucionando hacia una auténtica ontología ligera (*Schema.org*, s.d.).

Referencias

Berners-Lee, Tim; **Hendler, James**; **Lassila, Ora** (2001). *The semantic web*. *Scientific American*, May 17. *http://www.med.nyu.edu/research/pdf/mainim01-1484312.pdf*

DCMI Schema.org Alignment Task Group. *Schema.org Alignment*. *DCMI*, 2012. *http://wiki.dublincore.org/index.php/Schema.org_Alignment*

Schema.org (2012). "The type hierarchy". *Schema.org*. *http://schema.org/docs/full.html*

Schema.org (2012). "What is *Schema.org*? The type hierarchy". *Schema.org*. *http://schema.org*